Improving Automated Model Reconstruction across Phylogenetically Diverse Genome-Scale Metabolic Models.

José P. Faria[1], *Janaka N Edirisinghe*[1], *Samuel M.D. Seaver*[1], *Filipe Liu*[2], *Pamela Weisenhorn*[1], *James G. Jeffryes*[1], *Tian Gu*[1], *Qizh Zhang*[1], *and Christopher S. Henry*[1]
(1)Argonne National Laboratory, Argonne, IL, (2)Centre of Biological Engineering, University of Minho, Braga, Portugal

The Department of Energy Systems Biology Knowledgebase (KBase) is a platform designed to solve the grand challenges of Systems Biology. KBase has implemented bioinformatics tools that allow for multiple workflows including genome annotation, comparative genomics, and metabolic modeling. KBase now also includes a comprehensive database of over 100K reference genomes (approximately 5K complete genomes) from NCBI's RefSeq. We selected a phylogenetically diverse set of approximately 1000 genomes and constructed draft genome-scale metabolic models using the ModelSEED pipeline implemented in KBase. We used these 1000 genomes as a training set to improve the quality of models produced by the ModelSEED pipeline. First, we curated our mapping of RAST functional roles to biochemistry based on data mined from KEGG and published metabolic models; we corrected errors in our reaction reversibility assertions to improve overall model constraints; we applied a new method to predict auxotrophy across all 1000 genomes to predict ideal gapfilling media; we improved biomass formulations based on asserted presence or absence associated biosynthesis pathways; and we use proteins from KEGG and published metabolic models to improve and find genes to map to gapfilled reactions. We show how all of these corrections increase the number of gene associations, decrease the number of gapfilled reactions with no gene associations, and decrease the number of blocked reactions across all models.

We apply our improved modeling pipeline to construct and compare models of 5000 complete prokaryotic genomes. We show how biomass composition, auxotrophy, and pathway presence varies across these genomes along the phylogenetic tree. We also plot model quality across the phylogenetic tree, identify taxa where model quality is lower. We identify pathways and biomass components that are poorly characterized, particularly focusing on pathways with limited functionality using current biomass positions and objective functions in our metabolic models.

All draft models are available for download from KBase.